BMC Medical Genomics

RESEARCH

Open Access

Dynamic clustering of genomics cohorts beyond race, ethnicity—and ancestry

Hussein Mohsen[1,2,3*], Kim Blenman[4,5], Prashant S. Emani[6], Quaid Morris[1,7], Jian Carrot-Zhang[2] and Lajos Pusztai[4]

Abstract

Background Recent decades have witnessed a steady decrease in the use of race categories in genomic studies. While studies that still include race categories vary in goal and type, these categories already build on a history during which racial color lines have been enforced and adjusted in the service of social and political systems of power and disenfranchisement. For early modern classification systems, data collection was also considerably arbitrary and limited. Fixed, discrete classifications have limited the study of human genomic variation and disrupted widely spread genetic and phenotypic continuums across geographic scales. Relatedly, the use of broad and predefined classification schemes—e.g. continent-based—across traits can risk missing important trait-specific genomic signals.

Methods To address these issues, we introduce a dynamic approach to clustering human genomics cohorts based on genomic variation in trait-specific loci and without using a set of predefined categories. We tested the approach on whole-exome sequencing datasets in ten cancer types and partitioned them based on germline variants in cancer-relevant genes that could confer cancer type-specific disease predisposition.

Results Results demonstrate clustering patterns that transcend discrete continent-based categories across cancer types. Functional analysis based on cancer type-specific clusterings also captures the fundamental biological processes underlying cancer, differentiates between dynamic clusters on a functional level, and identifies novel potential drivers overlooked by a predefined continent-based clustering.

Conclusions Through a trait-based lens, the dynamic clustering approach reveals genomic patterns that transcend predefined classification categories. We propose that coupled with diverse data collection, new clustering approaches have the potential to draw a more complete portrait of genomic variation and to address, in parallel, technical and social aspects of its study.

Keywords Genetic variation, Cancer genomics, Classification, Ancestry, Ethnicity, Race

*Correspondence:
Hussein Mohsen
hussein.mohsen@utoronto.ca
[1] Computational and Systems Biology, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA
[2] Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA
[3] Terrence Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, ON M5S 3E1, Canada
[4] Breast Medical Oncology, School of Medicine, Yale University, New Haven, CT 06511, USA
[5] Computer Science, Yale University, New Haven, CT 06511, USA
[6] Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06511, USA
[7] Computational Biology and Medicine, Weill-Cornell Medical College, New York, NY 10065, USA

Background

In light of the growing availability of genomics datasets and the subsequent analyses of the complexities underlying both the human genome and genomic variation, the use of a fixed, predefined set of categories comes across as reductionist at best. In this paper, we utilize a qualitative (i.e. historical) and quantitative lens to highlight descriptive and applied problems underlying the use of broad discrete categories, including on a predefined continent-based level. Consequently, we present a dynamic trait-specific lens that clusters a genomics cohort's data based on variation in genomic loci associated with a trait under study without centering a predefined set of categories. We next demonstrate the utility of this lens in studying ten cancer types as examples of highly complex traits by identifying known and overlooked patterns on the clinical and functional genomic levels.

Historical perspective

The earliest scientific attempt to use race as a category to classify human beings dates back to the seventeenth century. In a 1684 essay titled, "A New Division of the Earth, According to the Different Species or Races of Men Who Inhabit it," French physician Francois Bernier categorized human beings into five types, the last of which, the Sámi people, he described using derogatory terms [1]. Swedish botanist Carl Linnaeus, dubbed as the founder of modern taxonomy, published decades later (1735) the first edition of *Systema Naturae*, in which he created a system with four categories instead. In the tenth edition (1758), he expanded the system and confounded physical with personality and social traits based on his interpretation of the humoral theory that links geography and climate to skin color and good and bad character [2]. In this work, Linnaeus loaded his classifications with prejudice and crafted a hierarchy placing *Homo sapiens europaeus*, a category he color-coded as white, on top, while using descriptions such as "harsh face," "careless," "stubborn," "lazy," "greedy," and "ruled by caprice" to describe *Homo sapiens americanus*, *afer* and *asiaticus*, color-coded respectively as red, black, and yellow [1, 3, 4]. Linnaeus also added a separate category he called *Homo sapiens monstrosus*, in which he mostly included humans with various birth defects and mythical "humans" such as giants from Patagonia [2–5].

Bernier and Linnaeus suggested different sets of categories and imbued their human classification systems with imagined hierarchical value judgement. Their imposition of a few, fixed, distinct and discrete categories reduced the complexity of human variation and shaped subsequent classification systems. In early systems of classification, descriptions often depended on the philosophical (and political) choices of the classifiers, technological limitations, and economic factors including trade routes [2, 6]. For instance, the routes involving Sweden and the Netherlands during Linnaeus' time shaped his considerably arbitrary choice to describe peoples of specific geographies but not others. Further, Linnaeus relied on anecdotal and written accounts of his students and of missionaries, mercantilists, travelers, and slave traders, and he did not travel himself outside of western Europe [2, 4].

Early classification systems were also venues for the use of emerging modern science to demarcate human difference in the service of power during times of colonial expansions and the Atlantic slave trade. Institutions and individuals exerted intentional efforts to create racial classification systems in modern science, which opened the door for racialized hypothesis generation. In a stark yet far from lone example, the Bordeaux Royal Academy of Science announced an essay contest in 1739 to study "the degeneration of Black hair and Black skin." The announcement was made a year after the regional assembly of Bordeaux endorsed the existence of enslaved Black people on French soil, and as recently described in "Who's Black and Why: A Hidden Chapter from the Eighteenth-Century Invention of Race," before members of the Bordeaux Academy decided to invest Academy prize money in the company that ran the French slave trade in the African continent, Compagnie Perpétuelle des Indes [7]. Another example of a (re)defined color line to benefit the interests of chattel slavery before its abolition in the United States is the introduction of "one drop laws," according to which a person was categorized Black if they had a known "trace of Black blood" in their ancestry. As a result, in the words of anthropologist Nina Jablonski, skin color was "no longer the necessary and sufficient criterion for race classification" [2] (for more examples, see [8, 9] and references listed in Box 1–1 of [10]).

Changing meanings of race categories continued to reflect and drive political and social transformations. Since the beginning of the first U.S. census in 1790, for example, racial groupings in the census have changed more than twenty times [1]. Notably, race categories and their meanings also vary across national borders [1, 10]. Biology and medicine are susceptible to societal and cultural influences, and scientists are engaged in a bi-directional process of being influenced by social and cultural concepts that co-shape interpretations of nature, and scientific interpretations that in turn influence social order [11, 12]. Scientific attempts to formulate classification systems of race continued in the nineteenth and twentieth centuries and were muddled, again, with confusion. In the words of anthropologist Fay-Cooper Cole during the opening of the "Conference on Racial Differences,"

which was held in 1928 at the National Academy in Washington, D.C. and attended by opponents and proponents of eugenics, the term "race" was "frequently used in three or four ways in the same article," and there existed "a great deal of confusion in the use of the word." [13] Eugenics, which propagated race science for more than six decades, declined during the 1930 s and 1940 s after it faced strong scrutiny and criticism within scientific circles and in response to Nazi eugenical horrors. Yet, the use of racial classification systems to study human variation continued [13]. The understanding of human variation has progressed since then, however, and further highlighted their unreliability.

Further limitations

Even with respect to skin color as a trait, the reduction of human variation into a small set of color-coded categories implies the existence of distinct skin color lines—when skin color is a continuum influenced by climate, genetics, and the intensity and seasonality of ultraviolet radiation [2]. Similar skin colors can also result from convergent adaptation in response to similar selective pressures, rather than from genetic relatedness [2, 14], and further analysis and data collection demonstrated the prevalence of continuous rather than discrete skin color distributions (e.g. [15, 16]). Further, comprehensive diverse genomics datasets have demonstrated (i) the complexity and prevalence of continuums on the genome level (e.g. [17]), and (ii) the sharp limitations of broadly predefined classification—be it at the level of socio-political categories such as race and ethnicity, or broad geographic ones such as continental ancestry (e.g. [18]).

While genetic ancestry is a concept that describes a partial relationship of a person with their genealogical history, it can still be subject to significant limitations as a classification criterion. First, there is no single criterion to define ancestry, and categories can take geographic (e.g. South Asian or Central American), geopolitical (e.g. Zambian or Italian), or cultural (e.g. Brahmin or Lemba) meanings [14]. Second, geographic ancestry categories might be muddled with imprecise conflation with race categories, and their descriptors can be distortive of time and space (e.g. references to Asian ancestry might exclude the entirety of or wide regions within West, Central, East or South Asia, and nationality-based categories might refer to ancestors during the period that preceded the very formation of respective countries). Third, continent-based categories reduce the high levels of genomic complexity within each continent and might inadvertently imply a nonexistent "purity" when communicating results. Fourth, separate categories impose discreteness on continuums between continents [14, 17, 19–21] that have long been connected by land (e.g. Asia and Europe, or today's Asia and Africa through the Sinai Peninsula), in technology, or both.

Further, single category assignments to individuals ignore the multitudes of personal belonging. While many individuals choose to affiliate with multiple groups for personal or cultural reasons [14], it is also highly common for individuals to have a genetic ancestry associated with multiple groups in sets of predefined categories or genetic panels (e.g. 97.3% of individuals are associated with a median of four ancestry categories in [18], in consistence with [22]). Importantly, this ancestry can be observed on the individual level and does not have to reflect population stratification. Further, significant amounts of genetic ancestry labeled as "Western Asian," for example, is present in samples with origins ranging from present-day Morocco to Mongolia, and from England to Ethiopia, that is, in Asia, Europe and Africa [18].

Dynamic, trait-specific germline clustering

Given the limitations of predefined classification systems, and recognizing the wide range of phenotypes and the complexity of genomic variation, we propose a dynamic approach that generates trait-specific clusterings of genomics cohorts. The approach builds on an earlier idea from an exchange between biological anthropologist Frank B. Livingstone and evolutionary biologist Theodosius Dobzhansky on generating clusterings on the gene(s) level (see [23] and Chapter 9 of [13]), and expands it in light of the wide advances in genomic data collection, measurement of genomic variation, and interpretation of the genomic basis of complex traits. The approach is also motivated by multiple factors. First, the genomic basis of different traits is encoded in different loci of the human genome, and the loci relevant for a single trait, ranging from one to many in number, cover only a small portion of the whole genome. Second, biological and physiological processes are shared among all humans. Third, especially when common germline polymorphisms are involved in part or in full in trait predisposition, the genomic variants are significantly shared across continental regions. Relatedly, a single nucleotide polymorphism (SNP) can be concurrently classified "rare" in multiple regions, and the distribution of classifications depends on available data [24].

Fourth, evolutionary forces might be acting on trait-specific genomic regions in parallel in distant geographies. Fifth, as predefined ancestral labels concurrently bear geopolitical, historical, and social meanings, their assignment to categories used to study genomic variation, and particularly predisposition to disease, opens the door for prolonging a history of stigmatizing entire communities [25].

Finally, the disruption of observed continental clines can overlook inter-continental patterns related to a trait of interest. This raises a core question on the goal of clustering cohorts: if two individuals in distant geographies have similar genomic markers and phenotypic expression corresponding to a trait, e.g. both are right-handed, should they be in the same cluster when studying the genomic basis of handedness, or separate ones? Should "populations" be determined based on a gene(s) (or trait) of interest, or the whole genome? What are the limitations of clustering based on the whole genome, in a fragmented and data-scarce setting, when a trait is affected by only a small subset of genomic regions? Further, it is also quantitatively well-established that selected features or clustering criteria strongly affect resulting "populations" (i.e. clusters or clines), and consequently reshape the starting point from which to discover—or miss—patterns and generate hypotheses [26, 27].

The dynamic clustering approach takes a different angle to classifying genomic variation by grouping individuals in a given cohort based on predisposition to a trait under study—herein a cancer type. An individual's membership to a cluster depends on their genomic sequence at a specific set of regions known to be associated with the trait. Number of clusters, which are de facto neutrally labelled, is determined according to the dataset and specific trait under study.

In cancer, germline (inherited) predisposition is mediated by deleterious mutations in several dozen high penetrance cancer-relevant genes and probably a combination of individually low penetrance variants. Different genes are associated with different degrees of risk, and with variable cancer-specificity [28]. Further, germline alterations require additional acquired (somatic) mutations for malignant transformation [29–31]. We hypothesize that clustering cancers based on their germline variants in cancer type-specific loci transcends predefined continent-based categories. We expect clusterings to vary across cancers—in terms of number of clusters and sample-cluster membership—in reflection of the varying levels of complexity underlying their genomic component. We also note that this dynamic (i.e. trait-specific) approach moves beyond the notion of local ancestry at a single locus as it can simultaneously consider any set of coding or non-coding regions associated with a (healthy or disease) trait, and it can scale to accommodate newly acquired knowledge on the genomic basis of the trait under study.

Methods

Genomic datasets

We used TCGA germline data from the breast invasive carcinoma (BRCA, n = 1072 samples after data processing), colon adenocarcinoma (COAD, 445), kidney renal clear cell carcinoma (KIRC, 514), liver hepatocellular carcinoma (LIHC, 360), lung adenocarcinoma (LUAD, 513), lung squamous cell carcinoma (LUSC, 503), ovarian serous cystadenocarcinoma (OV, 556), pancreatic adenocarcinoma (PAAD, 182), prostate adenocarcinoma (PRAD, 488), and rectum adenocarcinoma (READ, 164) studies. We used the MC3 somatic dataset [32] filtered according to the recommendations in [33] for potential driver identification and the ancestral labels obtained from the TCGAA Project (http://fcgportal.org/TCGAA/) [34].

SNP selection

For COSMIC-based SNP sets, we selected autosomal SNPs annotated as nonsynonymous, stop-gain, and stop-loss ClinVar database annotations [35]. For HFI subsets, the functional impact of missense germline variants within a cancer type's exome samples was determined using MetaSVM [36], SIFT [37], and MutationAssessor [38], and annotations by ClinVar (v20190305), when available. We considered a missense variant to have a high functional impact if it is categorized as Deleterious by MetaSVM or SIFT, High/Medium by MutationAssesor, or Pathogenic/Likely Pathogenic in ClinVar. We used MetaSVM, SIFT and MutationAssesor scores from the dbNSFP database (v35c via ANNOVAR [39]) which includes pre-calculated function impact scores for 75,931,005 human non-synonymous single-nucleotide variants [40]. We only included autosomal variants with GQ > 20 and alternative allele frequency > 20% and which met quality control measures described in Huang et al. [30]. The selected SNPs in each COSMIC and HFI variant set are listed in Supplementary Table 4.

Driver gene identification

Potential driver gene identification was performed on the MutSigCV [41] v1.3.4 server available at https://www.genepattern.org/modules/docs/MutSigCV. Genes with $q < 0.1$ were deemed statistically significant potential drivers.

Clinical variable analysis

Continuous and ordinal variable comparisons (i.e. for age, tumor grade, and tumor stage) were performed using the Wilcoxon rank-sum test in a one-vs-all configuration on clusters with > 5% of samples within a cohort and COSMIC or HFI setting, with Bonferroni correction and $p_{adj} < 0.05$ significance level.

Gene expression analysis

Differential expression analysis was performed using Moonlight v1.20 [42] (FDR < 0.05) at the gene program

level and edgeR v3.36 [43] (FDR < 0.05, $|Log_2 FC| > 2$) at the gene level.

Enrichment analysis

Enrichment analysis to identify pathways and biological processes was performed on g:Profiler available at https://biit.cs.ut.ee/gprofiler/, with entities having g:SCS threshold < 0.05 considered significant [44]. Visualization was done using ggplot2 [45], except for Fig. 6 generated using EnrichmentMap v.3.5.0 plugin [46] in Cytoscape v 3.10.3 [47].

MDS plotting and algorithmic clustering

We used PLINK v1.90 [48] available at https://zzz.bwh.harvard.edu/plink/ to calculate identity-by-state matrices (IBS) based on the allele values of the chosen variant set in each cohort, where sample pairs with closer genomic variant composition result in higher similarity values (--distance ibs in PLINK). IBS matrices were then used to generate input distance matrices (1 – IBS matrix) to classical multidimensional scaling (MDS), which is a dimensionality reduction algorithm that aims to preserve distances between samples in a lower dimensional space (cmdscale in the stats package in R: https://stat.ethz.ch/R-manual/R-devel/library/stats/html/00Index.html). For algorithmic clustering comparison (Fig. 3), DBSCAN clusters were identified using the dbscan package in R (https://cran.r-project.org/web/packages/dbscan/index.html) [49], and HClust (hclust) and K-Means (kmeans) using the stats package.

Results

Overview

To study predisposition to ten of the most common cancer types [50], we utilized germline data from the Cancer Genome Atlas (TCGA) [28, 30] and ancestral category values from the The Cancer Genetic Ancestry Atlas (TCGAA) [34]—which are based on comparisons with the 1000 Genomes [51], the Human Genome Diversity (HGDP) [52] and the International HapMap [53] Projects. In studied cohorts, TCGAA uses categories that refer to continental and sub-continental regions. For clarity, we hereafter refer to them collectively as continent-based.

We first generated the trait-specific clusterings of each cohort, namely BRCA, COAD, KIRC, LIHC, LUAD, LUSC, OV, PAAD, PRAD, and READ, using all nonsynonymous SNPs within different sets of COSMIC genes known to have germline association with each cancer type (Methods, Supplementary Table 3). Next, we focused on the SNP subset predicted to have high functional impact within each cancer type's samples as a basis for dynamically generating clusters. We then assessed the performance of three algorithmic clustering approaches (K-means, DBSCAN, and HClust) to identify generated clusters in multidimensional scaling (MDS) plots. Finally, we identified potentially overlooked somatic driver genes in each TCGA cohort based on dynamic clustering in comparison with drivers identified using a continent-based lens, and performed a functional genomic analysis to assess their biological and clinical importance in the context of cancer.

Beyond continent-based categories

Upon dynamically clustering based on trait-specific regions—herein cancer type-specific germline COSMIC genes, a number of visual patterns emerge. First, the number of clusters varies per cancer type (1–8 clusters), strongly reflecting known genomic heterogeneity in cancer [28, 30, 54] (Fig. 1). Second, and despite the relative lack of diversity in TCGA datasets, clusters transcend continent-based categories in all cancer types to include samples with "African," "East Asian," "European," and "Other" ancestral labels within clusters. Third, this pattern is also observed in colon and rectum cancers, known to be associated with high disparities in incidence and outcome [55, 56], with one and two clusters (Fig. 1b and 1j), respectively.

High-functional-impact compact clusters

Next, we selected subsets of SNP variants with high functional impact (HFI) on protein function within the COSMIC genes corresponding to each TCGA cohort as a basis for clustering (see Methods). This selection led to SNP subsets with n = 1 (PRAD) to 269 (BRCA). HFI-based dynamic clusters transcend continent-based categories and exhibit two notable patterns. First, the number of HFI-based clusters is higher, on average, than clusters based on all nonsynonymous SNPs of COSMIC genes (e.g. PAAD with two COSMIC-based vs four HFI-based clusters in Fig. 1h and Fig. 2a, respectively; HFI-based clusters for other cancer types in Supplementary Fig. 1). Second, subsections within select cancer types tend to be compact and often include samples with highly similar or identical HFI variant patterns due to the smaller number of loci used for clustering in these cohorts. Dots corresponding to distinct samples overlay each other,

Human aid improves algorithmic clustering

Algorithmic approaches can generate different clusterings of the same dataset, and their performance primarily depends on data distribution and cluster definition [27, 57]. To algorithmically identify clusters, we ran the K-means algorithm with a predefined k = 4 (number of selected TCGAA continent-based categories) and an "optimal" k chosen based on visual inspection of cancer

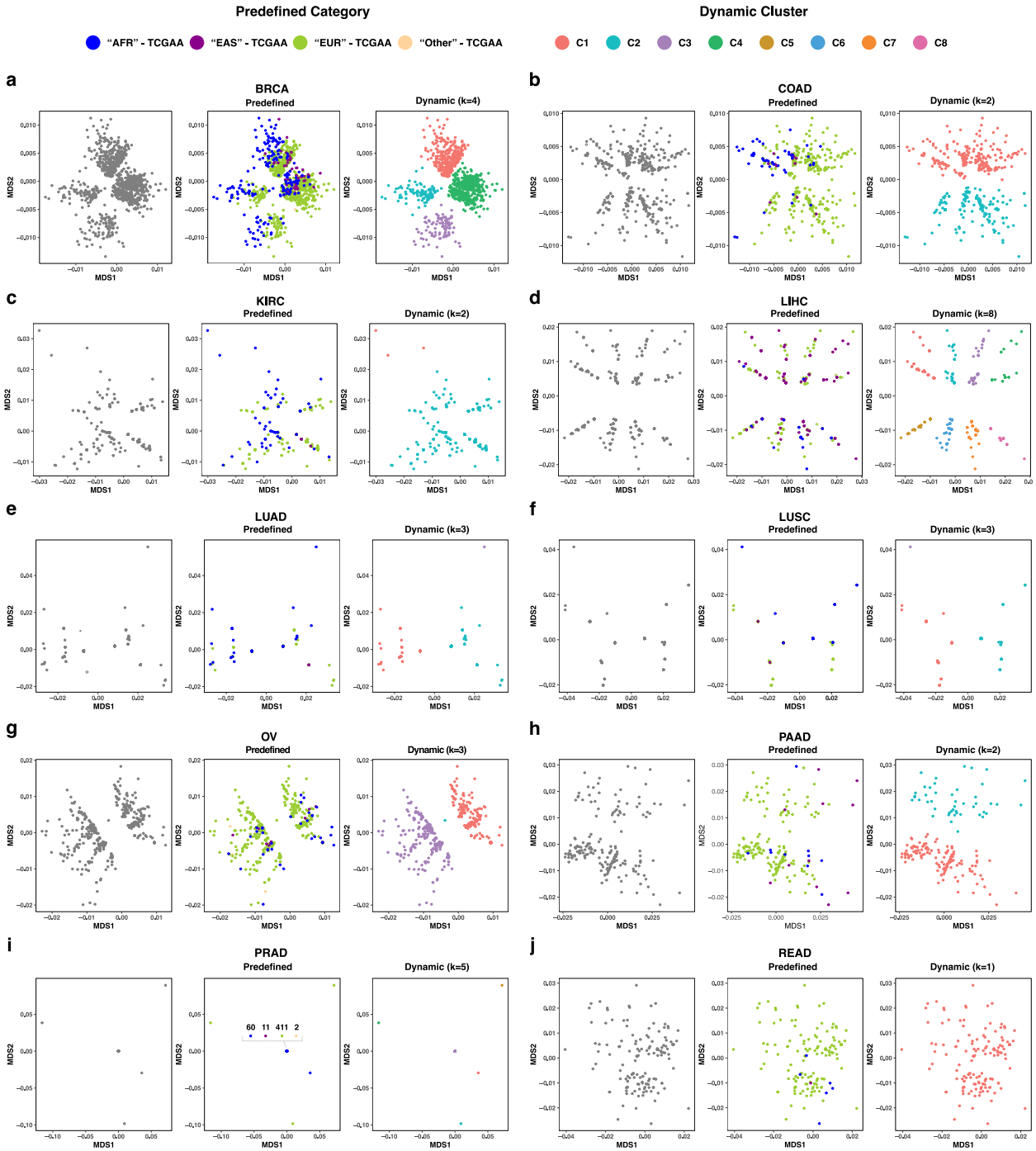

Fig. 1 Multidimensional scaling (MDS) plots of dynamically generated clusters for ten TCGA cancer cohorts. Cancer type-specific dynamic clusters transcend predefined continent-based categories. Dynamic cluster numbers (i.e. C1, C2, … C8) correspond to disjoint sample subsets within each cancer cohort. resulting in single dots each representing a subcluster (e.g. LUSC-HFI in Fig. 2b and KIRC-HFI in Supplementary Fig. 1).

type-specific plots. Given the varying number of clusters across cancer types, predefined-k clustering performed poorly across multiple cancer types (Fig. 3a-d). Similarly, dynamic-k K-means faced challenges in accurately identifying clusters across cohorts (e.g. LIHC-COSMIC, Fig. 3e; Supplementary Fig. 2).

Similarly, clustering results from two other algorithms, DBSCAN (Density-based Spatial Clustering of

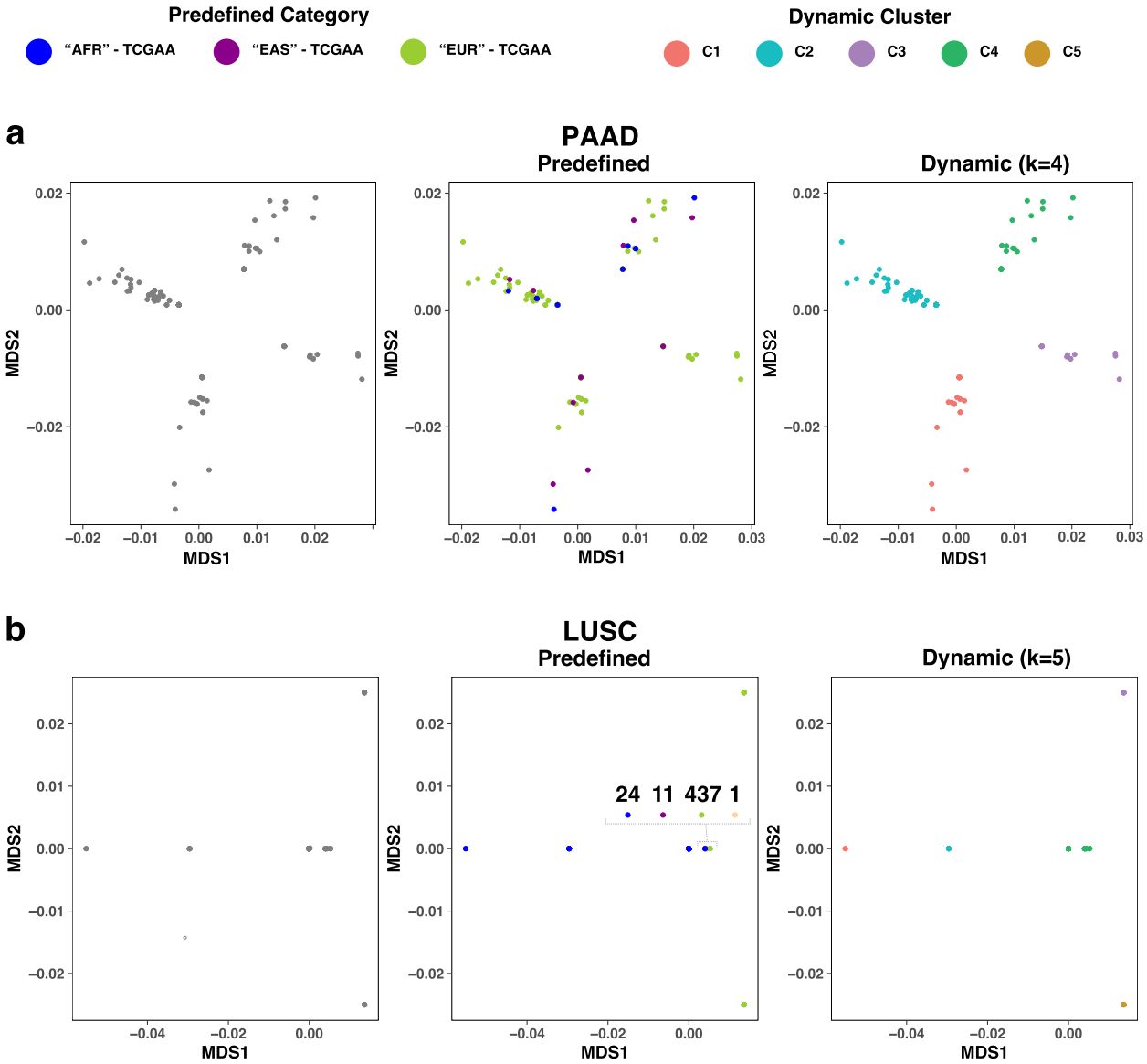

Fig. 2 MDS plots of dynamically generated clusters based on high-functional-impact (HFI) germline variant subsets. **a** PAAD results demonstrate a higher number of clusters in HFI-based results compared to ones based on all nonsynonymous variants in the COSMIC-based setting in Fig. 1h. **b** LUSC results demonstrate compact clusters with a high number of samples demonstrating similar or identical variation patterns in HFI subsets

Applications with Noise) and HClust, performed poorly in multiple cohorts. HClust, which stands for agglomerative hierarchical clustering—herein used with complete-linkage, is a bottom-up approach that starts with individual points as separate clusters and iteratively merges most similar clusters until a predefined number of clusters is met (e.g. k = 8 clusters for LIHC-COSMIC in Fig. 3f). Like K-means, HClust correctly identified only a subset (i.e. two) of the eight clusters in LIHC-COSMIC. DBSCAN, which is known to identify dense clusters and outliers in low-density regions, partitioned LIHC-COSMIC results into three clusters (Fig. 3g), a number decided algorithmically based on input parameters (see Methods): two large clusters, each roughly with four of the dynamic clusters, and a third cluster that includes distant samples the algorithm considered outliers (in red). Notably, DBSCAN faced more limitations with other cancer types such as READ and PAAD (Supplementary Fig. 4).

In sum, and while specific algorithms perform better than others at identifying clusters, algorithmic clustering does not seem to suffice for both COSMIC- and

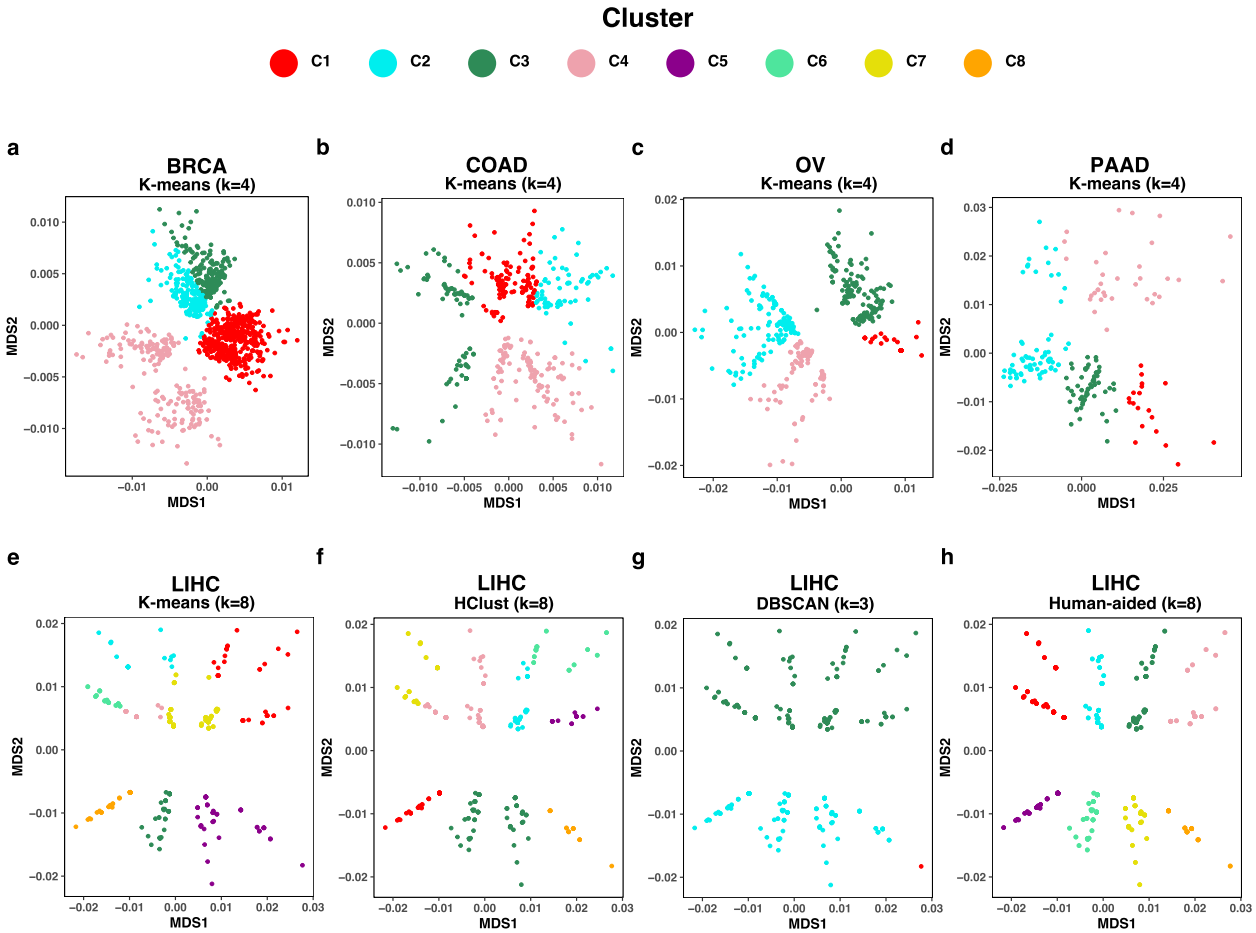

Fig. 3 Algorithmic and human-aided identification of dynamic clusters. K-means results with predefined-k = 4 fails to identify COSMIC-based clusters in (**a**) BRCA, (**b**) COAD, (**c**) OV, and (**d**) PAAD among other cancer types. Dynamic-k results also demonstrate the failure of (**e**) K-means, (**f**) HClust, and (**g**) DBSCAN to identify COSMIC-based clusters in LIHC, highlighting the need for (**h**) human-aid in cluster identification

HFI-based settings (Supplementary Figs. 2–7). As a result, human intervention to attain more precise results remains central (e.g. Fig. 3h). We also note that in certain instances, multiple "optimal" numbers of clusters in the same plot can exist, and the choice remains centered on experimental goals and the problem under study (e.g. the four dynamic clusters of BRCA-COSMIC in Fig. 1a being alternatively considered two larger diagonally-separated clusters in the same plot).

Dynamically-identified cancer drivers

We next explored the biological significance of clusters identified by the dynamic clustering approach. Particularly, we focused on identifying potential cancer type-specific somatic driver genes. We used MutSigCV to identify potential somatic drivers based on whole exome sequence data corresponding to each dynamic cluster generated on germline variant sets (COSMIC and HFI). COSMIC-based clusters yielded 98 potential drivers across cancer types, and HFI-based clusters 109 drivers. Both lists included a wide range of known drivers from a more comprehensive list by Bailey et al. that relied on multiple computational and experimental tools [58]. These include *KRAS*, *TP53*, *PIK3 CA*, *BRCA1*, *PTEN*, *CDH1*, *RB1*, *PTEN*, *FOXA1*, *SPOP*, and *VHL* (for full lists, see Supplementary Tables 1 and 2). COSMIC- and HFI-based lists also include 31 and 36 cancer type-specific novel potential drivers, respectively, which are overlooked if analysis is performed on continent-based clusters (Fig. 4a). Among these genes are known drivers listed in [58], including *APC*, *CBFB*, *B2M*, *CDKN2 A*, and *RPL5*.

We then investigated the functional importance of COSMIC- and HFI-based driver gene lists. Given their significant coverage of known drivers, enrichment analysis of both full lists point to known oncogenic pathways, including the majority of signaling pathways listed in Sanchez-Vega et al. [59]. These

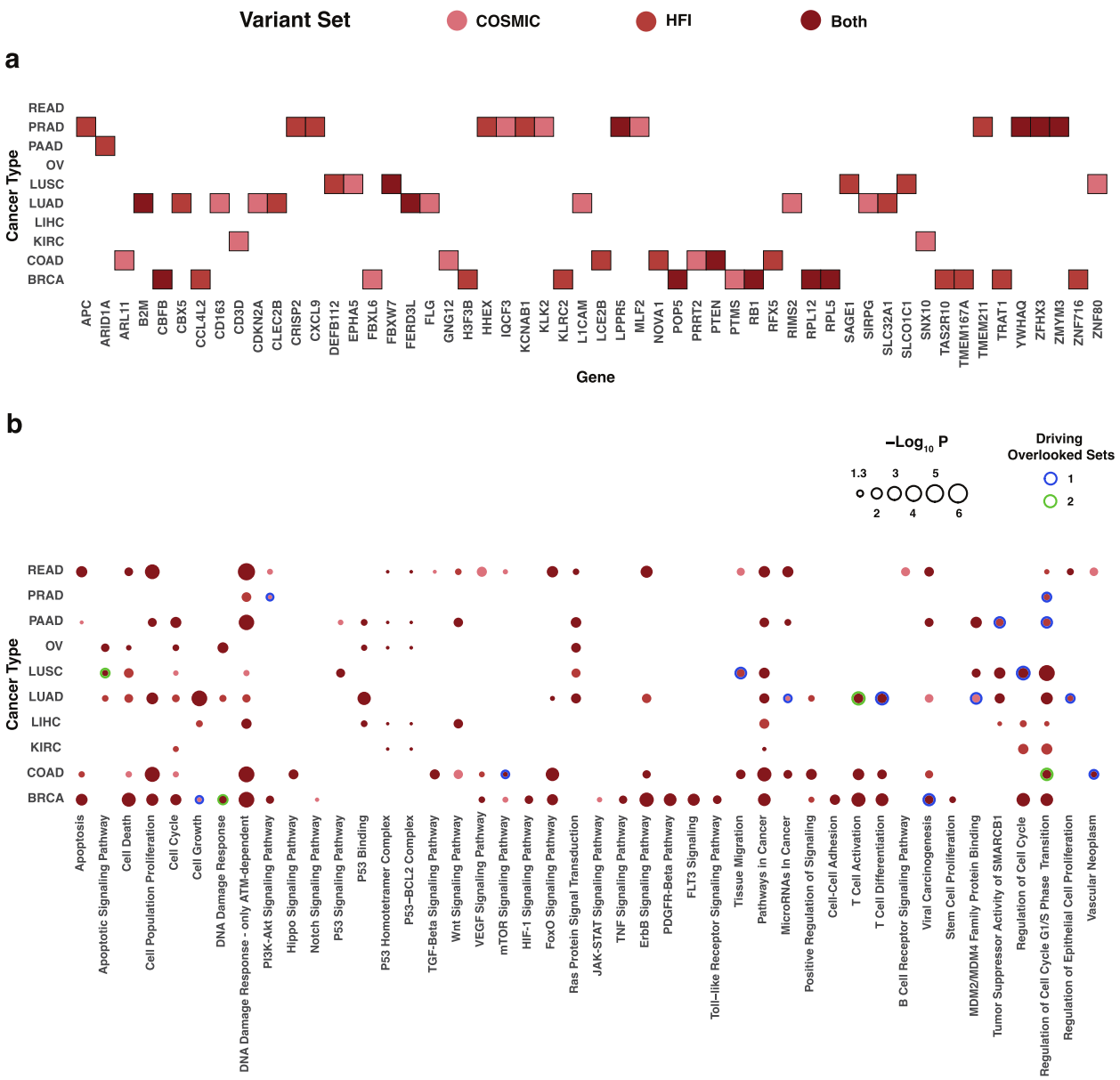

Fig. 4 Known and potential driver genes identified based on dynamic clustering. **a** Dynamic cluster-based genes overlooked by the continent-based scheme. Each of the listed genes was identified in at least one COSMIC- or HFI-based dynamic cluster and none of the clusters based on predefined continent-based categories. **b** Dynamic cluster-based drivers associate widely with known cancer pathways. Genes overlooked by the continent-based scheme drive a subset of these associations in one (blue border) or both (light green) settings centered on the COSMIC- and HFI-based variant sets

pathways include cell cycle alongside Hippo, Notch, PI3 K/Akt, TP53, TGFB and WNT signaling. Among other important pathways and processes are apoptosis, HIF-1, mTOR, TNF, JAK-STAT, VEGF, and FoxO signaling (Fig. 4b), and pathways named after several cancer types and other diseases and infections (e.g. Epstein-Barr and Kaposi sarcoma-associated herpesvirus virus infections). We compared enrichment results associated with the dynamically-generated lists with and without novel drivers overlooked by the continent-based scheme. The inclusion of novel genes allows dynamic clusters to refer to biological processes, pathways, and entities with strong effect on cancer etiology (blue and light green borders) such as apoptotic signaling in LUSC, DNA damage response in BRCA, mTOR signaling in COAD, and multiple terms pertaining to

cell adhesion, tissue migration, cell cycle, and cell proliferation across multiple cancer types.

Dynamic clusters are distinguishable by clinical and functional cancer signifiers

Dynamic clusters vary by the composition of their underlying germline variants. To investigate the biomedical significance of the resulting clusterings, we tested for the associations between each of the clusters, compared to its all its counterparts within a cancer cohort, with clinical variables and gene expression patterns in TCGA. In the LUAD-COSMIC setting, the age at the first pathologic diagnosis is lower in cluster 1 (C1) than 2 (mean = 64.4 and 66.6 years, respectively; Wilcoxon rank-sum test, $p_{adj} < 0.05$; Fig. 5a). In LIHC, C1 of the COSMIC setting shows a significant enrichment for high tumor grade samples, and C1 of the HFI setting for late tumor stage ones (Fig. 5b and 5c, respectively; Wilcoxon rank-sum test, $p_{adj} < 0.05$), with similar results that vary among clusters in LIHC and LUSC cohorts (Supplementary Fig. 8).

We then shifted attention to analyzing differential gene expression patterns across cancer types and settings. At the individual gene level, opposite expression levels are detected among clusters of the same cohort (Fig. 5d). Notably, such patterns correspond to genes with reported associations with tumorigenesis, patient survival, metastasis, and cell proliferation. These include *CSN2* [60], *NROB1* [61], and *NR1H4* [62] in both COSMIC and HFI settings in BRCA, and *MT1B* [63], *MUC1* [64], and *KLK11* [65] in LIHC-COSMIC. At the gene program level, we used Moonlight [42] to identify programs that are collectively differentially expressed in each cluster compared to matched "normal" samples. Within different cohorts, the magnitude and direction of expression mark significant differences among clusters. These

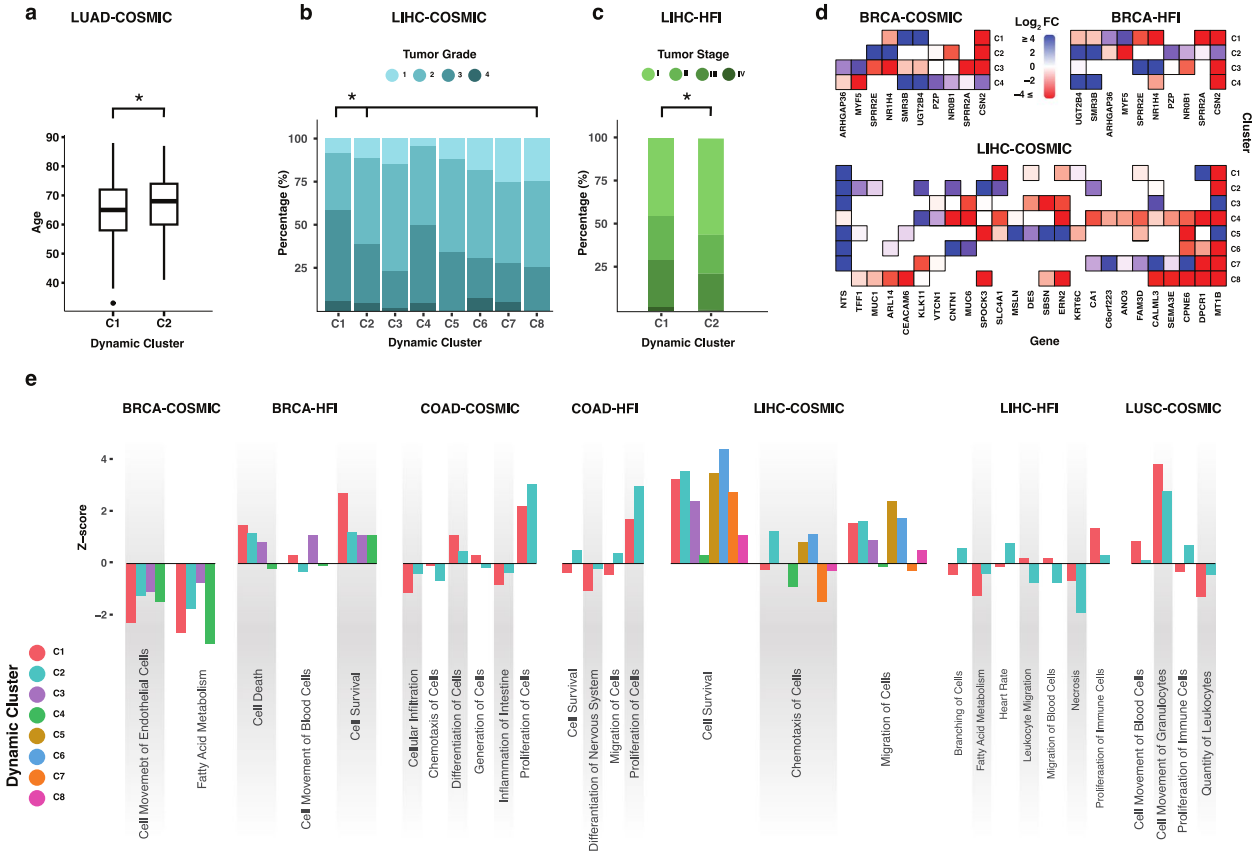

Fig. 5 Dynamic clusters across cancer types highlight clinical and functional associations. **a** Dynamic cluster 1 (C1) based on the COSMIC subset in LUAD (LUAD-COSMIC) shows statistically significant lower age cancer onset than that of the second cluster (C2; $p_{adj} < 0.05$). **b** C1 in LIHC-COSMIC includes samples with higher tumor grade compared to other clusters combined ($p_{adj} < 0.05$). Clusters C5 and C8 include no Grade 4 samples. **c** C1 shows more advanced tumor stage samples in LIHC-HFI compared to C2 ($p_{adj} < 0.05$). **d** Genes significantly expressed (FDR < 0.05) in opposite directions among clusters of BRCA-COSMIC, BRCA-HFI and LIHC-COSMIC highlight potential functional roles associated with different clusters. **e** Gene programs with known association to cancer are collectively expressed in different magnitudes and directions between dynamic clusters across cancer types and settings

include, among others (Fig. 5e; Supplementary Fig. 9), a negative expression (Z-score < 0) of cell death in only one out of four clusters in BRCA-HFI (C1), of migration of cells in only two out of eight in LIHC-COSMIC (C4 and C7), and of cell survival and migration of cells in one of the two clusters in COAD-HFI (C1); a considerably higher expression of proliferation of cells in one cluster in each of COAD-COSMIC and COAD-HFI (C2 in each); and a considerably lower expression of cell survival in LIHC-COSMIC (C4), quantity of leukocytes in LUSC-COSMIC (C2), and necrosis and fatty acid metabolism in LIHC-HFI (C1 and C2, respectively).

Dynamic clusters are distinguishable by non-cancer signifiers

In addition to cancer-focused gene programs, we investigated the biological significance of specific genes significantly expressed in only one dynamic cluster within each setting ($|\text{Log}_2\ \text{FC}| > 2$). Biological enrichment analysis of these gene sets revealed essential and non-cancer-focused biological processes that distinguish different clusters. These include, among others (Fig. 6, Supplementary Fig. 10), multiple processes related to neuronal response to stimulus in cluster 2 (C2) of COAD-COSMIC and neuron development in C1 of LIHC-HFI. Other clusters pertain to known associations between pathways, non-cancerous diseases, or infections with a specific tissue or cell type. These include the associations between lung squamous cell cancer and each of asthma [66] and type 1 diabetes mellitus [67] in C1 of LUSC-COSMIC and those between READ and the IL-17 pathway [68] in C2 of READ-HFI. Associations that closely pertain to cancer from an essential point of view include ones revolving around immune response in C2 of READ-HFI, as well as signaling and cell differentiation, which recurrently emerged from individual clusters within multiple tissues and cancer types (i.e. label "Across Cancer Types" in Fig. 6).

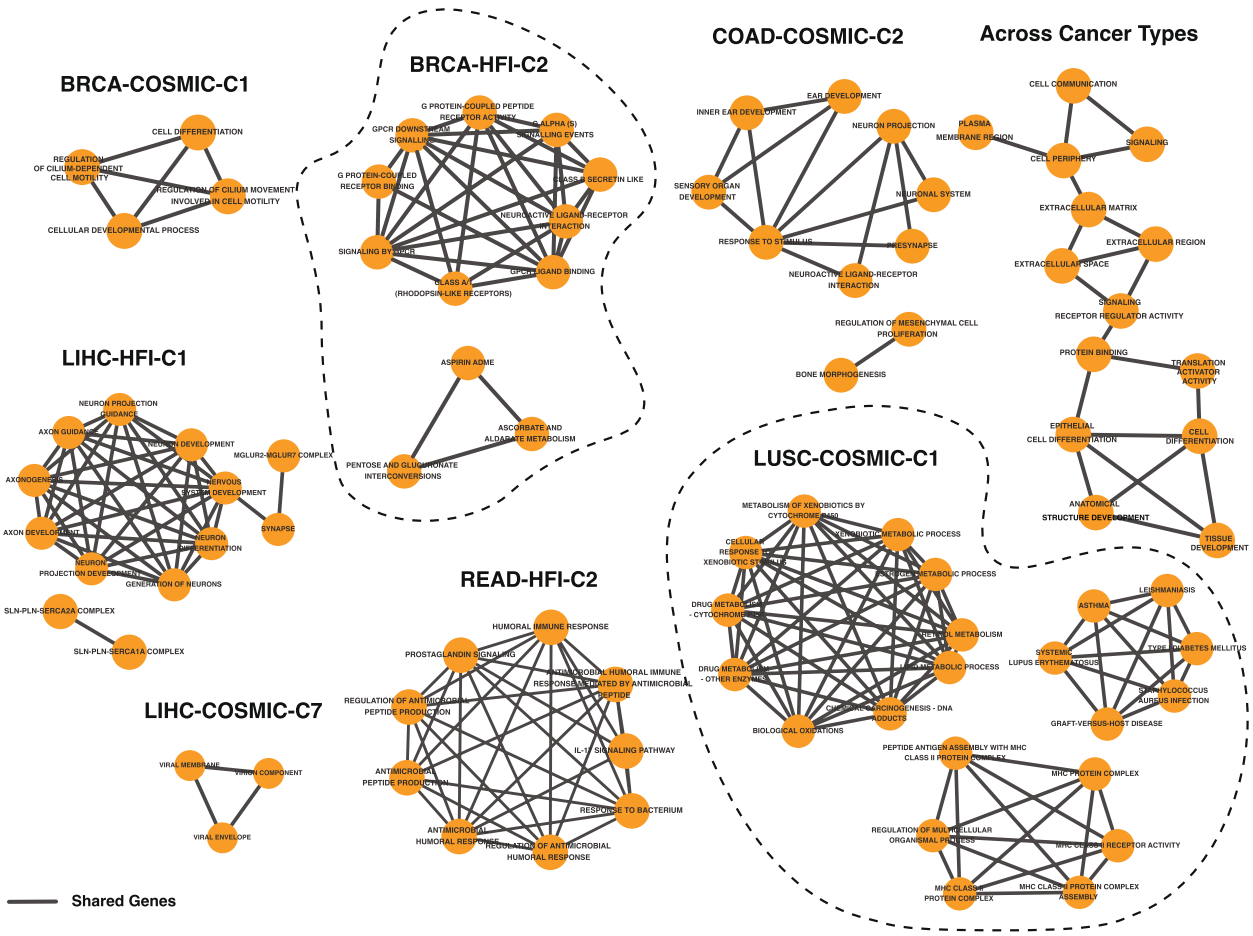

Fig. 6 Genes expressed in single clusters within each cancer type and setting highlight related biological and clinical associations beyond cancer (e.g. asthma and lupus in LUSC-COSMIC-C1 and neural development in LIHC-HFI-C1). Resulting associations share a subset of their underlying genes (i.e. edges), and a number of biological processes recurrently emerges across cancer types and settings (i.e. "Across Cancer Types," top-right)

Discussion

We introduced a dynamic approach to clustering human genomic variation that does not lock samples within predefined geography-based ancestry categories or average genomic patterns across the whole genome regardless of the trait under study. This approach also recognizes continuums and clusters when they exist at the trait-specific level (e.g. eight clusters for COSMIC-based LIHC vs one continuous cluster for READ in Fig. 1). When we apply the dynamic approach to germline data of the TCGA cancer cohorts, emerging clusters transcend predefined continent-based categories. When we examine the somatic mutational patterns in the resulting dynamic germline-based clusterings, we identify tens of known and potential cancer driver genes, many of which are overlooked by the continent-based scheme. Results based on trait-specific clusters also capture the fundamental biology associated with the hallmarks of cancer and associate clusters with clinical and biological signifiers [69].

The dynamic approach has broader implications for how human genomic variation is observed and classified. The use of racial classification systems in science has long been contested given their history that is rife with confusion, technological limitations, and enforcement in service of colonialism. Race, itself, is an idea rather than a discovery; an idea invented to impose systems of control and discrimination that continue to shape today's social realities [70]. Enforced color lines disrupted continuous clines of variation and often shifted to serve political goals rather than to describe patterns of variation. Further, race categories are usually collected to comply with civil rights reporting guidelines or for social and administrative purposes, but the racial categorization systems were not designed for genetic studies [2, 14]. Similarly, ethnicity categories can be centered on culture, social norms, religious beliefs, or language rather than genetic ancestry, and their use in genetic studies can lead to inaccurately reported results. Ethnicity categories are malleable concepts that can change in different times or circumstances irrespective of hereditary lines [14, 18].

A sharp decreasing trend in using race categories in genomic studies has been reported [71]. While genetic ancestry categories resemble, in their biological aspect, a direct reflection of a partial inheritance of genetic material across generational lines, they can also suffer from limitations in semantics, in space, and in time. Ancestry categories can be based on geographic, geopolitical, social, or cultural elements, and the process of imposing clear lines among predefined populations is rife with social and technical limitations. These limitations are heightened when categories are quite broad (e.g. continent-based), when individuals identify with more than one category, and when the labels might further open the door for enforced stigmatic associations of disease on entire communities [25]. As a result, it is generally advisable to cluster genomics cohorts only when justified by the research question rather than by default [10, 19], and to place social implications of the research at the heart of the design process rather than as an afterthought [72].

Given the compound nature of human genomic variation, a broader sampling of human genomic diversity—with the careful selection of categories during the data collection phase, if and as needed or obligated—remains highly central to more clearly understand the patterns of genomic variation (see Chapter 5 of [10] and [73]). In fact, it is through this type of diverse data collection efforts that continuous patterns of variation have been elucidated on a wider scale [17, 20]. Relatedly, alleles that increase the susceptibility to a disease can be present across multiple geographies. The notion of dynamic clustering can be carried over to genome-wide analyses as well. For example, in the case of methods that study one variant at a time—such as QTL identification and GWAS—it is possible to limit the SNP-based clustering to some neighborhood of each variant to consider how the local genetic relatedness of individuals can be accounted for in enrichment analysis.

While certain genomic patterns might be identified in a given dataset based on a given predefined model, this type of models is not necessarily the only route towards this type of identification. Equally importantly, as we demonstrate, other patterns can be missed when relying on broad categories or when considering complex traits with loci distributed across the genome. Broad stratification, whether driven by discriminatory legacies embedded in genetic practice or normalized experimental design and data collection, might obfuscate genomic patterns that transcend predefined categorical boundaries. Diverse datasets and new and existing quantitative approaches that can transcend predefined categories—by utilizing or being able to incorporate trait-specific regions, clines, estimated relatedness matrices (e.g. [74, 75]), principal components (e.g. [76–79]), or other chosen means—are hence crucial to approach different types of genomic studies (e.g. ones for gene discovery or other types of genomic studies described in [10]), to detect various patterns of genomic variation, to address theoretical and applied challenges (e.g. gene-environment interactions, pleiotropy, false positive control, sample size and statistical power considerations), and to leverage data generated using different technologies (e.g. deep sequencing and GWAS). These efforts have the potential to draw a more complete portrait of the genomic bases of traits all the while navigating the entangled relationships between science and society [11, 80–82].

Conclusions

Coupled with recent wide genomic data availability, existing, complex, and ever-changing human genomic variation engenders the need for multiple perspectives to studying genomic traits. We introduced a dynamic clustering approach that focuses on trait-based genetic similarity through a lens that centers both the technical and the social aspects of studying human genomic variation. We applied this approach to genomics cohorts corresponding to ten cancer types. Results demonstrated a varying number of germline clusters among cancer types, each of which transcending predefined, continent-based categories. Further analysis of these clusters captured known fundamental biology underlying cancer and identified potential cancer-related biomarkers that would be overlooked by a lens that is based on continent-based classification schemes.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12920-025-02154-z.

Supplementary Material 1.
Supplementary Material 2.
Supplementary Material 3.
Supplementary Material 4.
Supplementary Material 5.
Supplementary Material 6.
Supplementary Material 7.
Supplementary Material 8.
Supplementary Material 9.
Supplementary Material 10.
Supplementary Material 11.

Acknowledgements
No acknowledgements to include in this section.

Authors' contributions
Conceptual basis: HM. Methods and experimental design: HM, KB and LP. Experiment execution and data analysis: HM. Discussions: HM, KB, PE, QM, JCZ, and LP. First manuscript draft: HM and LP. Final manuscript: HM and LP with input from KB, PE, QM, and JCZ. All authors read and approved the final manuscript.

Funding
This study was supported by a Breast Cancer Research Foundation Investigator Award (BCRF-22–133) and a Susan Komen Leadership Grant (SAC220225) to LP.

Data availability
The results published here are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga. Controlled-access germline variants of TCGA cohorts were downloaded from the Genomic Data Commons (GDC, https://gdc.cancer.gov/about-data/publications/PanCanAtlas-Germline-AWG) of the National Cancer Institute (NCI) per Huang et al. [30]. TCGA variant and meta-data are available through the GDC portal at https://portal.gdc.cancer.gov/. Ancestry categories were obtained from TCGAA [34] available at http://fcgportal.org/TCGAA/. We chose to not analyze samples labeled "Native American [NA]" out of respect for Indigenous sovereignty. Clinical TCGA data was obtained from [83]. Gene expression data corrected for batch effect and study-specific bias were downloaded from RNAseqDB [84] at https://github.com/mskcc/RNAseqDB. Genes with germline associations at the tissue-specific level were downloaded from the COSMIC v90 [85] census list's 'Germline' column. Full gene lists ($n = 2$ to 11) are available in Supplementary Table 3.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 10 January 2025 Accepted: 6 May 2025
Published online: 15 May 2025

References

1. Roberts DE. Fatal invention: How Science, Politics, and Big Business Re-create Race in the Twenty-first Century. New York: New Press; 2011.
2. Jablonski NG. Skin color and race. Am J Phys Anthropol. 2021;175(2):437–47.
3. Marks J. Long shadow of Linnaeus's human taxonomy. Nature. 2007;447(7140):28–28.
4. Anemone RL. Race and Human Diversity: A Biocultural Approach. Oxford and New York: Routledge; 2019.
5. Sax, B., When Adam and Eve Were Monkeys: Anthropomorphism, zoomorphism, and other ways of looking at animals, in The Routledge companion to animal-human history, H. Kean and P. Howell, Editors. 2018, Routledge/Taylor & Francis Group,: London ; New York.
6. HoSang, D.M., On Racial Speculation and Racial Science: A Response to Shiao et al. Sociological Theory, 2014. 32(3).
7. HL Gates AS Curran W Black Why?: A Hidden Chapter from the Eighteenth-century Invention of Race. 2022 Cambridge The Belknap Press of Harvard University Press Massachusetts
8. Curran AS. The Anatomy of Blackness: Science & Slavery in an Age of Enlightenment. Baltimore: Johns Hopkins University Press; 2011.
9. Hogarth RA. Medicalizing Blackness: Making Racial Difference in the Atlantic World, 1780–1840. Chapel Hill: The University of North Carolina Press; 2017.
10. National Academies of Sciences, Engineering, Medicine, Using Population Descriptors in Genetics and Genomics Research: A New Framework for an Evolving Field. Washington. DC: The National Academies Press; 2023.
11. Reardon J. The Human Genome Diversity Project: A case study in Coproduction. Soc Stud Sci. 2001;31(3):357–88.
12. Jasanoff, S., The Idiom of Co-Production, in States of Knowledge: The Co-Production of Science and the Social Order, S. Jasanoff, Editor. 2004, Routledge. p. 1–12.
13. Yudell M. Race Unmasked: Biology and Race in the Twentieth Century. New York: Columbia University Press; 2014.
14. Race, Ethnicity, and Genetics Working Group, The use of racial, ethnic, and ancestral categories in human genetics research. Am J Hum Genet, 2005. 77(4): p. 519–32.
15. Crawford, N.G., et al., Loci associated with skin pigmentation identified in African populations. Science, 2017. 358(6365): p. eaan8433.
16. Jacobs LC, et al. Comprehensive candidate gene study highlights UGT1A and BNC2 as new genes determining continuous skin color variation in Europeans. Hum Genet. 2013;132(2):147–58.
17. Wojcik GL, et al. Genetic analyses of diverse populations improves discovery for complex traits. Nature. 2019;570(7762):514–8.

18. Baker, J.L., C.N. Rotimi, and D. Shriner, Human ancestry correlates with language and reveals that race is not an objective genomic classifier. Scientific Reports, 2017. 7(1).
19. Sohail, M., A. Izarraras-Gomez, and D. Ortega-Del Vecchyo, Populations, Traits, and Their Spatial Structure in Humans. Genome Biology and Evolution, 2021. 13(12).
20. Lewis ACF, et al. Getting genetic ancestry right for science and society. Science. 2022;376(6590):250–2.
21. Serre D, Pääbo S. Evidence for gradients of human genetic diversity within and among continents. Genome Res. 2004;14(9):1679–85.
22. Shriner, D., et al., Genome-wide genotype and sequence-based reconstruction of the 140,000 year history of modern human ancestry. Scientific Reports, 2014. 4(1).
23. Livingstone, F.B. and T. Dobzhansky, On the Non-Existence of Human Races. Current Anthropology, 1962. 3(3).
24. Cotter, D.J., et al., A rarefaction approach for measuring population differences in rare and common variation. Genetics, 2023. 224(2).
25. Kader, F.Đ., Lan N.; Lee, Matthew; Chin, Matthew K.; Kwon, Simona C.; Yi, Stella S., Disaggregating Race/Ethnicity Data Categories: Criticisms, Dangers, And Opposing Viewpoints. Health Affairs Forefront, 2022.
26. Alelyani, S.T., Jiliang; Liu, Huan, Feature Selection for Clustering: A Review, in Data Clustering: Algorithms and Applications, C.C.R. Aggarwal, Chandan K., Editor. 2014, Chapman and Hall/CRC: New York, NY.
27. Ultsch, A. and J. Lötsch, The Fundamental Clustering and Projection Suite (FCPS): A Dataset Collection to Test the Performance of Clustering and Data Projection Algorithms. Data, 2020. 5(1).
28. ICGC-TCCA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. Nature. 2020;578(7793):82–93.
29. Qing, T., et al., Germline variant burden in cancer genes correlates with age at diagnosis and somatic mutation burden. Nature Communications, 2020. 11(1).
30. Huang, K.L., et al., Pathogenic Germline Variants in 10,389 Adult Cancers. Cell, 2018. 173(2): p. 355–370 e14.
31. Carter H, et al. Interaction Landscape of Inherited Polymorphisms with Somatic Events in Cancer. Cancer Discov. 2017;7(4):410–23.
32. Ellrott, K., et al., Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. Cell Syst, 2018. 6(3): p. 271–281 e7.
33. Bailey MH, et al. Retrospective evaluation of whole exome and genome mutation calls in 746 cancer samples. Nat Commun. 2020;11(1):4748.
34. Yuan, J., et al., Integrated Analysis of Genetic Ancestry and Genomic Alterations across Cancers. Cancer Cell, 2018. 34(4): p. 549–560 e9.
35. Landrum MJ, et al. ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res. 2016;44(D1):D862–8.
36. Kim S, et al. Meta-analytic support vector machine for integrating multiple omics data. BioData Min. 2017;10:2.
37. Sim, N.L., et al., SIFT web server: predicting effects of amino acid substitutions on proteins. Nucleic Acids Res, 2012. 40(Web Server issue): p. W452–7.
38. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res. 2011;39(17):e118–e118.
39. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38(16):e164–e164.
40. Liu, X., et al., dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. Human Mutation, 2016. 37(3): p. 235–241.
41. Lawrence MS, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013;499(7457):214–8.
42. Colaprico A, et al. Interpreting pathways to discover cancer driver genes with Moonlight. Nat Commun. 2020;11(1):69.
43. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139–40.
44. Raudvere U, et al. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Res. 2019;47(W1):W191–8.
45. H Wickman ggplot2: Elegant Graphics for Data Analysis. 2016 New York Springer-Verlag NY
46. Reimand J, et al. Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA. Cytoscape and EnrichmentMap Nature Protocols. 2019;14(2):482–517.
47. Shannon P, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13(11):2498–504.
48. Purcell S, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559–75.
49. Hahsler, M.P., Matthew; Arya, Sunil; Mount, David, dbscan: Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and Related Algorithms. 2022.
50. NCI SEER. Cancer Stat Facts: Cancer of Any Site. 2022; Available from: https://seer.cancer.gov/statfacts/html/all.html.
51. The 1000 Genomes Project Consortium, A global reference for human genetic variation. Nature, 2015. 526(7571): p. 68–74.
52. Cavalli-Sforza, L.L., et al., Call for a worldwide survey of human genetic diversity: a vanishing opportunity for the Human Genome Project. Genomics, 1991. 11(2).
53. International HapMap Consortium. The International HapMap Project. Nature. 2003;426(6968):789–96.
54. Pon JR, Marra MA. Driver and passenger mutations in cancer. Annu Rev Pathol. 2015;10:25–50.
55. Bailey CE, et al. Increasing disparities in the age-related incidences of colon and rectal cancers in the United States, 1975–2010. JAMA Surg. 2015;150(1):17–22.
56. Myer PA, et al. The Genomics of Colorectal Cancer in Populations with African and European Ancestry. Cancer Discov. 2022;12(5):1282–93.
57. Xu D, Tian Y. A Comprehensive Survey of Clustering Algorithms. Annals of Data Science. 2015;2(2):165–93.
58. Bailey MH, et al. Comprehensive Characterization of Cancer Driver Genes and Mutations. Cell. 2018;173(2):371-385.e18.
59. Sanchez-Vega F, et al. Oncogenic Signaling Pathways in The Cancer Genome Atlas. Cell. 2018;173(2):321-337.e10.
60. Bach K, et al. Time-resolved single-cell analysis of Brca1 associated mammary tumourigenesis reveals aberrant differentiation of luminal progenitors. Nat Commun. 2021;12(1):1502.
61. Zhang H, et al. The prognostic value of the orphan nuclear receptor DAX-1 (NROB1) in node-negative breast cancer. Anticancer Res. 2011;31(2):443–9.
62. Journe F, et al. Association between farnesoid X receptor expression and cell proliferation in estrogen receptor-positive luminal-like breast cancer from postmenopausal patients. Breast Cancer Res Treat. 2009;115(3):523–35.
63. Xu P, et al. Copy number variation of metallothionein 1 (MT1) associates with MT1X isoform expression and the overall survival of hepatocellular carcinoma patients in Guangxi. Gene Reports. 2024;34: 101889.
64. Yuan SF, et al. Expression of MUC1 and its significance in hepatocellular and cholangiocarcinoma tissue. World J Gastroenterol. 2005;11(30):4661–6.
65. Ke H, et al. Serum protein biomarkers for HCC risk prediction in HIV/HBV co-infected people: a clinical proteomic study using mass spectrometry. Front Immunol. 2023;14:1282469.
66. Rosenberger A, et al. Asthma and lung cancer risk: a systematic investigation by the International Lung Cancer Consortium. Carcinogenesis. 2011;33(3):587–97.
67. Yu, X., et al., Causal relationship between diabetes mellitus and lung cancer: a two-sample Mendelian randomization and mediation analysis. Frontiers in Genetics, 2024. 15.
68. Razi S, et al. IL-17 and colorectal cancer: From carcinogenesis to treatment. Cytokine. 2019;116:7–12.
69. Hanahan D, Robert A. Weinberg, Hallmarks of Cancer: The Next Generation. Cell. 2011;144(5):646–74.
70. McLean S-A. Isolation by Distance and the Problem of the Twenty-First Century. Hum Biol. 2019;91(2):81–94.
71. Byeon YJJ, et al. Evolving use of ancestry, ethnicity, and race in genetics research—A survey spanning seven decades. The American Journal of Human Genetics. 2021;108(12):2215–23.
72. Reardon, J., Human Population Genomics and the Dilemma of Difference, in Reframing Rights: Bioconstitutionalism in the Genetic Age, S. Jasanoff, Editor. 2011, MIT Press: Cambridge, Massachusetts; London, Englad. p. 217–238.

73. Oni-Orisan A, et al. Embracing Genetic Diversity to Improve Black Health. N Engl J Med. 2021;384(12):1163–7.
74. Kang HM, et al. Variance component model to account for sample structure in genome-wide association studies. Nat Genet. 2010;42(4):348–54.
75. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nat Genet. 2012;44(7):821–4.
76. Jiang L, et al. A resource-efficient tool for mixed model association analysis of large-scale data. Nat Genet. 2019;51(12):1749–55.
77. Mbatchou J, et al. Computationally efficient whole-genome regression for quantitative and binary traits. Nat Genet. 2021;53(7):1097–103.
78. Zhao, H., et al., A practical approach to adjusting for population stratification in genome-wide association studies: principal components and propensity scores (PCAPS). Statistical Applications in Genetics and Molecular Biology, 2018. 17(6).
79. Zhou W, et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nat Genet. 2018;50(9):1335–41.
80. Yudell M, et al. Taking race out of human genetics. Science. 2016;351(6273):564–5.
81. Thorp, H.H., Time to look in the mirror, in Science. 2020.
82. Nobles, M.W., Chad; Wonkam, Ambroise; Wathuti, Elizabeth, Science must overcome its racist legacy: Nature's guest editors speak, in Nature. 2022.
83. Liu J, et al. An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. Cell. 2018;173(2):400-416.e11.
84. Wang Q, et al. Unifying cancer and normal RNA sequencing data from different sources. Scientific Data. 2018;5(1): 180061.
85. Tate JG, et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Res. 2019;47(D1):D941–7.

Publisher's Note